A survey on multimodal large language models

Year: 2023 | *Citations: 924* | *Authors: Shukang Yin, Chaoyou Fu, Sirui Zhao*

Abstract

ABSTRACT Recently, the multimodal large language model (MLLM) represented by GPT-4V has been a new rising research hotspot, which uses powerful large language models (LLMs) as a brain to perform multimodal tasks. The surprising emergent capabilities of the MLLM, such as writing stories based on images and optical character recognition–free math reasoning, are rare in traditional multimodal methods, suggesting a potential path to artificial general intelligence. To this end, both academia and industry have endeavored to develop MLLMs that can compete with or even outperform GPT-4V, pushing the limit of research at a surprising speed. In this paper, we aim to trace and summarize the recent progress of MLLMs. First, we present the basic formulation of the MLLM and delineate its related concepts, including architecture, training strategy and data, as well as evaluation. Then, we introduce research topics about how MLLMs can be extended to support more granularity, modalities, languages and scenarios. We continue with multimodal hallucination and extended techniques, including multimodal in-context learning, multimodal chain of thought and LLM-aided visual reasoning. To conclude the paper, we discuss existing challenges and point out promising research directions.